An Anomaly Detection Method for Weighted Data Based on Feature Association Analysis

Jiayao Li
School of Software, Shanxi Agricultural University, China
lijiayao@sxau.edu.cn

Rui Wang
Department of Computer Engineering, Shanxi Polytechnic College, China
17696066288@163.com

Abstract: *In recent years, weighted data is appearing more and more frequent in many applications, but the existence of anomalies decreases the accuracy of data-based operations, thus, it is necessary to detect anomalies to improve the data quality. However, the existing anomaly detection methods for weighted data only consider the Weighted Frequent Itemsets (WFIs) or Weighted Rare Itemsets (WRIs) separately, which causes their detection accuracy is seriously dependent on the preset minimal weighted support (min_wsup) value. To address these issues, we propose an anomaly detection method for weighted data on the basis of feature association analysis, namely ADWD, it accurately detects the anomalies under different min_wsup values through fully considering both WFIs and WRIs. ADWD first deletes infrequent 1-itemses during constructing Weighted Frequent Itemset-based Tree (WFI-Tree), thus decreasing time overhead on the inquiry of extensible itemsets; And then, ADWD defines three deviation metrics through comprehensively considering possible influencing factors to calculate transaction's abnormal score. Finally, the transactions whose abnormal score in top-rank are judged as anomalies. Extensive experiments on three datasets verify that the proposed ADWD method can more accurately detect anomalies from weighted data within less time usage, as well as has good scalability.*

Keywords: *Anomaly detection, feature association analysis, weighted data.*

Received June 25, 2023; accepted December 5, 2023
https://doi.org/10.34028/iajit/21/1/11

1. Introduction

On this big data era, data plays an important role in every aspect of real life. With the use of large-scale data, researchers propose massive clustering methods [21], classification methods [24, 26], prediction methods [9], etc., thereby providing more robust support for the life and production. However, the collected data usually contains abnormal data due to acquisition equipment failure, network transmission error and so on. The existence of abnormal data will produce relatively large effect on data-based operations, therefore, how to accurately detect abnormal data from large-scale data to improve data quality is an urgent problem to solve.

In many practical scenarios, each data element (i.e., feature, itemset) has a weight value representing its importance (such as popularity, rating, etc.,), the larger weight values represent the higher importance of the feature or greater popularity. For example, the webpages have different clicks and the larger clicks represent this webpage is more popular by users. In recent years, the weighted data (that is, each feature is accompanied by a weight value representing its importance) is more common in real life, and it has gradually become a main form of data. Although existing distance-based anomaly detection methods [13, 20], clustering-based anomaly detection methods [3, 22], density-based anomaly detection methods [2, 14 27], feature association-based anomaly detection methods [4, 5, 6], and deep learning-based anomaly detection methods [15, 18] can accurately detect the contained abnormal data, but most methods default that each feature is equally important. That is, most methods do not consider the influence of feature weights on the results of anomaly detection, which causes these methods cannot be effectively used to detect abnormal data from weighted data.

The weight-sensitive anomaly detection methods (such as AvgDiff [16], WFP-Outlier [25], MWRPM-Outlier [8], WMFP-Outlier [7]) detected anomalies through considering the association of each feature as well as the weight value of features, they achieved higher detection accuracy than other categories of anomaly detection methods. However, they had the following problems [17, 23]:

1) The AvgDiff method only uses weight value as a means to improve the detection efficiency, it is not suitable for processing weighted data.
2) WFP-Outlier method does not consider more factors that cause transactions to be abnormal data in the abnormal detection stage, which results in it having lower detection accuracy.
3) Although the WMFP-Outlier method improves the detection accuracy of WFP-Outlier through considering more influencing factors, but its detection accuracy declines seriously when the minimum weighted support threshold (recorded as

min_wsup) is set larger.

4) Although the MWRPM-Outlier method solves the lower detection accuracy problem of WMFP-Outlier under large min_wsup via mining minimal weighted rare itemsets, but its detection accuracy is poorer under smaller min_wsup. These problems of existing anomaly detection methods prompt us to design a more balanced anomaly detection algorithm to make its detection accuracy less dependent on the min_wsup value.

Through fully analyzing the WMFP-Outlier and MWRPM-Outlier methods, we find that the sharply drop of detection accuracy under different min_wsup is that these two methods just simply mine Weighted Frequent Itemsets (WFIs) or Weighted Rare Itemsets (WRIs), which is easily affected by the scale of mined weighted itemsets. That is, the sharply decrease number of mined Maximum Weighted Frequent Itemsets (MWFIs) under large min_wsup value in WMFP-Outlier method makes it having significant drop of detection accuracy; The dramatically decrease number of Minimum Weighted Rare Itemsets (MWRIs) under larger min_wsup value in MWRPM-Outlier method makes it having significant drop of detection accuracy. Aiming at the problems existing in WMFP-Outlier and MWRPM-Outlier methods for weighted data, this paper designs and implements an effective feature association analysis-based anomaly detection method for weighted data, namely ADWD. The contributions are concluded as:

1) We comprehensively mine the MWFIs and MWRIs than only one category of weighted itemsets to reduce the problem of low detection accuracy caused by the reduction number of mined weighted itemsets.
2) Based on the mined MWFIs and MWRIs, we design three deviation metrics through considering more possible influences to calculate the transaction's abnormal score.
3) On the basis of designed deviation metrics, we propose an anomaly detection method called ADWD for weighted data to seek for anomalies from weighted data.
4) We perform extensive experiments to test whether the proposed ADWD method can achieve better detection efficiency as well as better scalability on three publicly available datasets, and the experimental results confirm that ADWD can obtain higher precision, recall, F1-measure and accuracy within less time consumption, as well as has better scalability.

The remainder of this paper can be organized as follows. Section 2 reviews some related works. In section 3, we provide some preliminaries. Section 4 introduces anomaly detection method based on the feature association analysis. In section 5, we use three datasets to perform massive experiments to verify the efficiency of ADWD method. Finally, we conclude the contributions and discuss the future direction.

2. Related Works on Feature Association Analysis-based Anomaly Detection Methods

Feature association analysis-based anomaly detection is a category of anomaly detection method through mining the associations between features. This category of anomaly detection method first uses data mining technology to mine itemsets with frequently appearance or rarely appearance, and then designs several deviation metrics to measure the abnormality of transactions, where the transactions whose abnormal score in top rank are judged as anomalies [5, 6]. In this category of anomaly detection methods, the time efficiency and detection accuracy are two major aspects to consider, where the feature mining phase aims at solving the long time consumption problem on the mining of associated features and the abnormal detection phase aims at solving the low detection accuracy problem [7, 8].

To solve the problem of heavy time consumption, Giacometti and Soulet [10] proposed a Frequent Pattern Outlier Factor (FPOF) to discover anomalies from static datasets without mining all frequent itemsets. Although the time cost of FPOF has been greatly reduced, but its detection accuracy was not very ideal. In order to give more interpretation of outliers, Rasheed and Alhajj [19] proposed a periodic feature-based method to seek for the anomalies from data streams, it used a suffix tree as the underlying data structure and repeatedly measured its outlier degree using mined periodic features, which resulted it having certain advantages in processing time as well as having certain flexibility. Compared with frequent feature-based outlier detection method, the rare features indicate lower appearing frequency, which is more appropriate with the definition of outliers. Based on this idea, Hemalatha *et al.* [12] proposed a Minimum Infrequent Pattern-based Outlier Detection (MIFPOD) to seek for anomalies from data streams, where the anomaly degree of each transaction was measured by three designed deviation factors. MIFPOD method has competitive detection efficiency when processing large min_sup values, but the situation is vise under small min_sup values due to the small scale of mined minimum rare itemsets.

In recent years, three anomaly detection methods [7, 8, 25] were proposed to detect anomalies in the weighted data based on the mining of associated features, where WFP-Outlier [25] and WMFP-Outlier [7] were based on the WFIs and MWRPM-Outlier [8] was based on the WRIs. These three methods only considered the WFIs or WRIs singly, which caused their detection accuracy was seriously depending on the setting of min_wsup. To solve this problem, it is required to comprehensively adopt the WFIs and WRIs to improve the detection accuracy and thus making the anomaly detection not so relying on the setting to min_wsup value.

3. Preliminaries

The weighted data is very similar to that of traditional data, while the difference is that each itemset (i.e., feature of weighted data) is accompanied by a *weight* value representing its importance that stored in the weight table. That is, wtable={wei(I_1), wei(I_2), …, wei(I_n)}, where I_n represents the nth itemset and wei(I_n) represents the weight value of I_n. When analyzing the associations of features, the min_wsup is used to measure whether I_n appears frequently. If I_n appears frequent (that is, wei(I_n)≥min_wsup), then I_n is a WFI; Otherwise, I_n is a WRI. Table 1 shows an example of weighted data.

Table 1. A specific weighted data.

TID	Transactions	TID	Transactions
T_1	{I_a, I_b, I_c, I_e}	T_2	{I_b, I_c, I_d, I_e}
T_3	{I_a, I_b, I_d}	T_4	{I_a, I_b, I_d, I_f}
T_5	{I_a, I_b, I_c, I_d, I_e}	T_6	{I_b, I_d, I_e, I_f}
Itemset	**weight**	**Itemset**	**weight**
I_a	0.6	I_b	0.8
I_c	0.9	I_d	0.7
I_e	0.3	I_f	0.1

- Transaction Weight (TW): for each transaction T_i, its *transaction weight* value is the ratio of the sum *weight* of contained itemsets to its length (recorded as *len*(T_i)), which is shown in Equation (1).

$$TW(T_i) = \frac{\sum_{I_m \in T_i} wei(I_m)}{len(T_i)} \quad (1)$$

For the example in Table 1, the *TW* value of T_1 is *TW*(T_1)=(0.6+0.8+0.9+0.3)/4=0.65; The *TW* value of T_2 *TW*(T_2)=(0.8+0.9+0.7+0.3)/4=0.675; The *TW* value of T_3 is *TW*(T_3)=(0.6+0.8+0.7)/3=0.7; The *TW* value of T_4 is *TW*(T_4)=(0.6+0.8+0.7+0.1)/4=0.55; The *TW* value of T_5 is *TW*(T_5)=(0.6+0.8+0.9+0.7+0.3)/5=0.66; The *TW* value of T_6 is *TW*(T_6)=(0.8+0.7+ 0.3+0.1)/4=0.475.

- Weight Support (WS): for each itemset I_m, its weight support is the sum of transaction weight in which contains I_m, which is shown in Equation (2).

$$WS(I_m) = \sum_{I_m \in T_i} TW(T_i) \quad (2)$$

For the example in Table 1, the *WS* value of I_a is WS(I_a)=0.65+0.7+0.55+0.66=2.56; The *WS* value of I_b is WS(I_b)=0.65+0.675+0.7+0.55+0.66+0.475=3.71; The *WS* value of I_c is WS(I_c)=0.65+0.675+0.66=1.985; The *WS* value of I_d is WS(I_d)=0.675+0.7+0.55+0.66+ 0.475=3.06; The *WS* value of I_e is WS(I_e)=0.65+0.675+ 0.66+0.475=2.46; The WS value of I_f is WS(I_f)=0.55+ 0.475=1.025.

- Minimum Weighted Rare Itemset (MWRI): for an itemset I_m, if wei(I_m)<min_wsup and there is no subset of I_m (recorded as I_n) making wei(I_n)<min_wsup, then, I_m is a MWRI.

For the example in Table 1, when the min_wsup is set to 2, because wei(I_a)=2.56>2, wei(I_b)=3.71>2, wei(I_c)=1.985<2, wei(I_d)=3.06>2, wei(I_e)= 2.27>2, wei(I_f)=1.025<2, and no subset of {I_c} and {I_f} is a MWRI, then {I_c} and {I_f} are MWRIs.

- Maximum Weighted Frequent Itemset (MWFI): for an itemset I_m, if wei(I_m)≥min_wsup and there is no superset of I_m (recorded as I_n) making wei(I_n)≥min_wsup, then, I_m is a MWFI.

For the example in Table 1, when the min_wsup is set to 2, because wei (I_a, I_b)=2.56>2, wei (I_a, I_b, I_d)=1.91<2 and wei (I_a, I_b, I_e)=1.31<2, that is, the weight value of the supersets of {I_a, I_b} is less than min_wsup, then, itemset {I_a, I_b} is a MWFI.

4. Anomaly Detection for Weighted Data

Similar to traditional feature analysis-based anomaly detection methods, the feature analysis-based anomaly detection method for weighted data detects the anomalies using two phases with the consideration of *weight* value, including feature analysis phase and anomaly detection phase. Compared with other anomaly detection methods for weighted data, MWRPM-Outlier method [8] is more competitive in terms of detection accuracy and time efficiency when processing large min_wsup values, therefore, it is considered as the main referenced method.

The mining of MWRIs in MWRPM-Outlier is very similar to that of Apriori method [19], it constructs a matrix to record the *TW* value, thereby reducing the scanning times of features and thus reducing the time cost. However, MWRPM-Outlier has the following problems in the mining phase:

1) The calculation of WS value for each feature needs to scan the constructed matrix structure for one time, it is very time consuming
2) The MWRIs mining process does not delete the WRIs, which causes these itemsets also participate in the scanning operations, therefore, some meaningless time is additional added
3) The MWRIs mining process does not arrange the features, which causes the time cost on the determination of different transactions whether containing the features is very long. In addition, MWRPM-Outlier only considers the mined *MWRIs* to detect anomalies, which leads to the low detection accuracy problem under small min_wsup value. All these problems prompt us to revise the MWRPM-Outlier method in two phases, thereby improving its detection efficiency.

4.1. The Mining of MWRIs and MWFIs

Compared with Apriori-based method [19], the FP-Growth-based method [12] is an efficient category of data mining method, thus, the FP-Growth method is adopted in the mining of MWRIs and MWFIs. The mining process needs to scan the transactions to calculate the *WS* value, it is very time consuming. However, once the feature (itemset, denoted as $\{I_a, I_b\}$) and the feature (denoted as $\{I_a, I_c\}$) are appearing in the same transactions, then the *WS* value of $\{I_a, I_b\}$ and $\{I_a, I_c\}$ are equal, which causes the repeatedly calculation of *WS* values is extra. For this reason, we adopt the idea of MWRPM-Outlier method to discover the different parts of itemsets, thereby reducing the time cost on the calculation of *WS* values. For two itemsets $\{I_a, I_b\}$ and $\{I_a,I_c\}$, their itemset different part is defined as: $Dif(I_{abc})=T_{ab}-T_{ac}$, where T_{ab} means that itemset $\{I_a, I_b\}$ is appearing in transaction T, T_{ac} means that itemset $\{I_a, I_c\}$ is appearing in transaction T, while $(T_{ab}-T_{ac})$ means that the transaction contains $\{I_a, I_b\}$ but not contains $\{I_a, I_c\}$. Therefore, $T_{abc}=T_{ab}\cap T_{ac}=T_{ab}-Dif(I_{abc})$, which results in the calculation of $WS(I_a, I_b, I_c)$ is shown in Equation (3).

$$WS(I_{a,}I_{b,}I_c) = \sum_{T_i\in T_{abc}} TW(T_i) = \sum_{T_i\in T_{ab}} TW(T_i) - \sum_{T_i\in Dif(I_{abc})} TW(T_i) = WS(I_{a,}I_{b,}I_c) - \sum_{T_i\in Dif(I_{abc})} TW(T_i) \qquad (3)$$

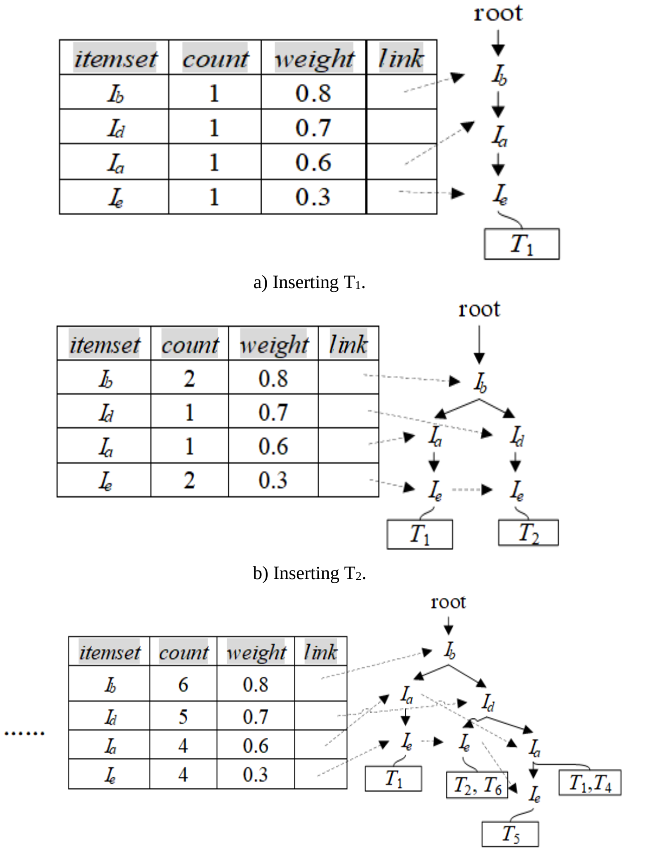
In the mining process of MWRIs and MWFIs based on the FP-Growth method, once two itemsets appearing in the same transactions (it is very convenient to judge because FP-Growth-based mining process needs to continuously scan the sub-trees to determine current itemset appears in which transaction and thus calculating the WS value), then, it is only to calculate one itemset's WS value. Because the scan of sub-trees is necessary, thus, the different parts of itemset are get easily, which leads to the time cost is not heavily.

And then, we introduce the mining process of MWRIs and MWFIs step-by-step with an example shown in Table 1, where min_wsup value is also set to 2.

- *Step* 1. Scan the data samples (aka, itemsets) in the transactions to calculate the WS value, and then discard the weighted rare 1-itemsets, while the weighted frequent 1-patterns are arranged by their decrease appearing times.

For the example shown in Table 1, the WS value of contained 1-itemsets is $WS(I_a)=2.56>2$, $WS(I_b)=3.71>2$, $WS(I_c)=1.985<2$, $WS(I_d)=3.06>2$, $WS(I_e)=2.46>2$, $WS(I_f)=1.025<2$, therefore, weighted rare 1-itemsets $\{I_c\}$ and $\{I_f\}$ are *MWRIs* and they need to be discarded to add into the following expanding operation. Because the appearing times for $\{I_a\}$, $\{I_b\}$, $\{I_d\}$ and $\{I_e\}$ are 4, 6, 5 and 3, respectively, thus, the inserting sequence is adjusted to $I_b \rightarrow I_d \rightarrow I_a \rightarrow I_e$.

- *Step* 2. Construct the Weighted Frequent Itemset Tree (WFI-Tree) for the weighted frequent 1-itemsets based on the decrease appearing times, the construction process is shown in Figure 1.
- *Step* 3. Mine the MWRIs and MWFIs from the leaf node to root node based on the constructed WFI-Tree, which is similar to that of FP-Growth method, while the difference is that the weight value of each feature needs to be considered.



itemset	count	weight	link
I_b	1	0.8	
I_d	1	0.7	
I_a	1	0.6	
I_e	1	0.3	

a) Inserting T1.

itemset	count	weight	link
I_b	2	0.8	
I_d	1	0.7	
I_a	1	0.6	
I_e	2	0.3	

b) Inserting T2.

itemset	count	weight	link
I_b	6	0.8	
I_d	5	0.7	
I_a	4	0.6	
I_e	4	0.3	

c) Inserting T6.

Figure 1. The construction of WFI-Tree.

For weighted frequent 1-itemset $\{I_e\}$, it can be extended to $\{I_e, I_a\}$, $\{I_e, I_b\}$ and $\{I_e, I_d\}$ through traversing the WFI-Tree. For $\{I_e\}$, it is appearing in T_1, T_2, T_5 and T_6: for $\{I_a\}$, it is appearing in T_1, T_3, T_4 and T_5, thus, $WS(I_e, I_a)=WS(I_e)-TW(T_2)-TW(T_6)=2.46-0.675-0.475=1.31<2$, $\{I_e, I_a\}$ cannot be further extended and it is a *MWRI*; For $\{I_b\}$, it is appearing in T_1, T_2, T_3, T_4, T_5 and T_6, thus, $WS(I_e, I_b)=WS(I_e)=2.46>2$, $\{I_e, I_b\}$ can be further extended; For $\{I_d\}$, it is appearing in T_2, T_3, T_4, T_5 and T_6, thus, $WS(I_e, I_d)=WS(I_e)-TS(T_1)=2.46-0.65=1.81<2$, itemset $\{I_e, I_d\}$ cannot be further extended and it is a *MWRI*. Because the weighted frequent 2-itemset extended by $\{I_e\}$ is only $\{I_e, I_d\}$, thus, it cannot be extended to longer itemset and it is a MWFI.

For weighted frequent 1-itemset $\{I_a\}$, it can be extended to $\{I_a, I_b\}$ and $\{I_a, I_d\}$ through traversing the WFI-Tree. For $\{I_a\}$, it is appearing in T_1, T_3, T_4 and T_5; For $\{I_b\}$, it is appearing in T_1, T_2, T_3, T_4, T_5 and T_6, thus, $WS(I_a, I_b)=WS(I_a)=2.56>2$, $\{I_a, I_b\}$ can be further extended; For $\{I_d\}$, it is appearing in T_2, T_3, T_4, T_5 and T_6, thus, $WS(I_a, I_d)=WS(I_a)-TS(T_1)=2.56-0.65=1.91<2$, $\{I_a, I_d\}$ cannot be further extended and it is a *MWRI*.

Because the weighted frequent 2-itemset extended by $\{I_a\}$ is only $\{I_a, I_b\}$, thus, it cannot be extended to longer itemset and it is a MWFI.

For weighted frequent 1-itemset $\{I_d\}$, it can be only extended to $\{I_d, I_b\}$ through traversing the WFI-Tree. For $\{I_d\}$, it is appearing in T_2, T_3, T_4, T_5 and T_6; For $\{I_b\}$, it is appearing in T_1, T_2, T_3, T_4, T_5 and T_6, thus, $WS(I_d, I_b)=WS(I_d)=3.06>2$. Because the weighted frequent 2-itemset extended by $\{I_d\}$ is only $\{I_d, I_b\}$, thus, it cannot be extended to longer itemset and it is a *MWFI*.

The specific mining operations of *MWRIs* and *MWFIs* can be concluded in Algorithm (1).

Algorithm 1: MWI-Mine

```
Input: Weighted Data (WD), min_wsup
Output: MWRIs, MWFIs
foreach (transactions in WD)
{
  foreach (1-item in transactions)
  {
    scan the transactions to find weighted frequent 1-item $\{I_a\}$
    {
      if ($WS(I_a)$<min_wsup)
        MWRIs ← $\{I_a\}$
      else
        use $\{I_a\}$ to construct WFI-Tree
    }
  }
}
k=2
foreach (itemset in WFI-Tree)
{
  {
    track k-itemset $\{I_a, I_b, \ldots, I_n\}$ in WFI-Tree
    if ($WS(I_a, I_b, \ldots, I_n)$<min_wsup)
      if (no subset of $\{I_a, I_b, \ldots, I_n\}$ is MWRI)
        MWRIs ← $\{I_a, I_b, \ldots, I_n\}$
    else
      if (no superset of $\{I_a, I_b, \ldots, I_n\}$ is MWFI)
        MWFIs ← $\{I_a, I_b, \ldots, I_n\}$
  }
  k++
}
return MWRIs and MWFIs
```

4.2. The Design of Deviation Metrics

In the mining phase, the mined MWRIs and MWFIs are used for the calculation of deviation factors of each transaction in the weighted data, where the design of deviation metrics is very critical to the anomaly detection. However, the existing feature-based anomaly detection methods only consider the influencing factors of MWRIs or MWFIs separately, but not fully consider both these two kinds of associated itemsets on the influencing of anomaly determination. To overcome this problem, we design three deviation metrics through considering following factors to improve the detection accuracy.

- Factor 1: the weight support and length of MWFIs. The large weight support value of MWFIs indicates these MWFIs appearing more frequent or these MWFIs are more important, which causes the transaction that contains this kind of MWFIs less likely to be an anomaly. In addition, the longer length of MWFIs indicates more WFIs are contained as the subsets of MWFIs, which leads to the transaction that contains longer length of MWFIs less like an anomaly. These two factors have a negative effect to the determination of anomaly.
- Factor 2: the length of similar parts between MWFIs and transaction. The longer length of the similar parts indicates that most itemsets in the transaction are WFIs, thus, this transaction is less like an anomaly.
- Factor 3: the weight support and length of MWRIs. Because MWRI has a positive effect to the determination of abnormal transaction, thus, the small weight support value of MWRI in the transaction (which means the MWRI appearing more rarely or having less importance) and short length of MWRI (which means more MWIs can be extended by this MWRI) will cause the transaction more abnormal.
- Factor 4: the number of contained MWRIs. Because MWRI indicates the itemset that appears rarely in the transaction or has less importance, which results in more abnormal of transactions caused by the large number of contained MWFIs, that is, this factor has a positive effect to the determination of anomaly.

- Deviation metric based on *MWFIs* ($DM\_MWFI(T_i)$): for the *MWFI* $\{X\}$, its length is $len(X)$ and its *weight support* is $WS(X)$, the similar part between *MWFI* and transaction T_i is $\{Y\}$, then, $DM\_MWFI(T_i)$ is defined in Equation (4).

$$DM\_MWFI(T_i) = \sum_{X\subseteq T_i} \left[\left(WS(X) - \min\ \_wsup\ \right) * 2^{len(X)}\right] + \frac{\sum_{Y\subseteq(X\cap T_i)} len(Y)}{len(T_i)} \tag{4}$$

- Deviation metric based on *MWRIs* ($DM\_MWRI(T_i)$): for the *MWRI* $\{A\}$, its length is $len(A)$ and its *weight support* is $WS(A)$, the number of contained *MWRIs* in transaction T_i is $num(A)$, then, $DM\_MWRI(T_i)$ is defined in Equation (5).

$$DM_{MWRI(T_i)} = \sum_{A\subseteq T_i} \left[\left(min\ \_wsup - WS(A)\right) * 2^{len(T_i)-len(A)}\right] + num(A) \tag{5}$$

- *Final deviation metric* ($FDM(T_i)$): $FDM(T_i)$ is a *comprehensive* factor that fully consider the influences of *DM_MWFI* and *DM_MWRI*, it is defined in Equation (6).

$$FDM(T_i) = DM\_MWRI(T_i) - DM\_MWFI(T_i) \tag{6}$$

4.3. The Details of ADWD Method

Based on the designed deviation metrics, the abnormal score of each transaction is calculated. And then, the transactions whose *FDM* value in top *k* are judged as anomalies. The specific process of ADWD is shown in

Algorithm (2).

Algorithm 2: ADWD

```
Input: Weighted Data (WD), min_wsup, k
Output: Anomalies
call Algorithm 1
DM_MWFI($T_i$)=0, DM_MWRI($T_i$)=0, FDM($T_i$)=0
foreach ($T_i$ in WD)
{
  foreach (MWFI {X} in $T_i$)
  {
    calculate DM_MWFI($T_i$)
  }
  foreach (MWRI {A} in $T_i$)
  {
    calculate DM_MWRI($T_i$)
  }
  calculate FDM($T_i$)
}
sort $T_i$ by their descending FDM($T_i$) value
Anomalies ← top k $T_i$
return Anomalies
```

As is shown in Algorithm (2), the MWFIs and MWRIs in the weighted data are mined with Algorithm 1 firstly, and then the DM_MWFI(T_i), DM_MWRI(T_i) and FDM(T_i) are set to 0 for initialization. For every transaction T_i in the weighted data, three deviation metrics are calculated based on Equations (4), (5), and (6), separately. Finally, all transactions are sorted according to the descending FDM(T_i) value, and the transactions whose FDM(T_i) value in top k rank are output as anomalies.

5. Experimental Analysis

In order to evaluate the detection capability of the proposed ADWD method, we carried out extensive experiments to answer the following three Research Questions (RQs).

- RQ 1: whether or not the proposed ADWD method can achieve higher detection accuracy compared with other state-of-the-art anomaly detection methods?
- RQ2: compared with state-of-the-art anomaly detection methods, can the proposed ADWD method consume shorter time?
- RQ3: does the proposed ADWD method can be effectively used in high dimensional datasets or large-scale datasets?

5.1. Setup of Experiments

1) Datasets: the datasets [1] used in the experiments include Lymphography, Wisconsin Breast Cancer Data (WBCD) and ForestCover, and the details are shown in Table 2. These three datasets used in the experiment are numerical in nature and each of them does not suffer from missing values. According to the rule provided in the datasets, the transactions in minority class are regarded as anomalies. Specifically, Lymphography dataset contains four classes, and the transactions in classes 2 and 4 are considered as anomalies; In the WBCD dataset, the transactions belonging to class 4 are considered as anomalies; *ForestCover* dataset has 54 features and the transactions in class 4 are considered as anomalies. Because these datasets do not provide *weight* value, thus, we randomly generate *weight* ranged from (0.0, 1.0) for each transaction to simulate the weighted environment.

Table 2. Characteristics of the used datasets.

Datasets	**Transactions**	**Dimensions**	**Anomalies**
Lymphography	148	18	6
WBCD	683	10	10
ForestCover	286048	10	2747

2) Implementation and environment: the run of all experiments is on an Inter(R) Core (TM) i7-10700 CPU, and the software environment is python 3.6.
3) Evaluation metrics: to measure the effectiveness of our proposed ADWD method in anomaly detection, we use recall, precision, F1-measure and accuracy metrics. We calculate these metrics as follows:

- Precision: the percentage of correctly detected anomalies amongst all detected anomalies by the method. *precision*=*TP*/ (*TP*+*FP*)
- Recall: the percentage of correctly detected anomalies amongst all anomalies by the method. *recall*=*TP*/ (*TP*+*FN*)
- F1-measure: The harmonic mean of precision and recall. *F*1-*measure*=2**precision***recall* / (*precision* + *recall*)
- Accuracy: the percentage of correctly detected *anomalies* and normal transactions amongst all transactions by the method. *accuracy*=(*TP*+*TN*)/ (*TP*+*FP*+*TN*+*FN*)

In these evaluation metrics, True Positive (TP) is the number of anomalies that were correctly identified by the method; False Positive (FP) is the number of normal transactions that were incorrectly identified as anomalies by the method; False Negative (FN*)* is the number of anomalies that were not correctly identified by the method; and True Negative (TN) is the number of normal transactions that were correctly identified by the method.

4) Compared methods: to evaluate the detection ability of the proposed ADWD method, the MWRPM-Outlier [8], WMFP-Outlier [7], WFP-Outlier [25], Adaptive-KD [1] and LODA [11] are compared in the experiments.

5.2. Answer to RQ1

To answer RQ1, we conduct extensive experiments on

[1] http://odds.cs.stonybrook.edu/

three publicly available datasets to test the detection accuracy, where different min_wsup values are used in the experiments. The experimental results are shown in Tables 3, 4, and 5.

Table 3. Detection accuracy on dataset Lymphography.

Metrics	Methods *min_wsup*	MWRPM-Outlier	WMFP-Outlier	WFP-Outlier	Adaptive-KD	LODA	ADWD
precision	29.60	50.00%	83.33%	83.33%	50.00%	50.00%	83.33%
	35.52	50.00%	83.33%	66.67%	50.00%	50.00%	83.33%
	41.44	66.67%	66.67%	66.67%	50.00%	50.00%	83.33%
	47.36	66.67%	66.67%	50.00%	50.00%	50.00%	83.33%
	53.28	83.33%	66.67%	50.00%	50.00%	50.00%	83.33%
recall	29.60	42.86%	71.43%	71.43%	42.86%	37.50%	83.33%
	35.52	42.86%	71.43%	57.14%	42.86%	37.50%	83.33%
	41.44	57.14%	66.67%	57.14%	42.86%	37.50%	83.33%
	47.36	57.14%	57.14%	42.86%	42.86%	37.50%	83.33%
	53.28	71.43%	57.14%	42.86%	42.86%	37.50%	83.33%
F1-measure	29.60	46.16%	76.92%	76.92%	46.16%	42.86%	83.33%
	35.52	46.16%	76.92%	61.54%	46.16%	42.86%	83.33%
	41.44	61.54%	66.67%	61.54%	46.16%	42.86%	83.33%
	47.36	61.54%	61.54%	46.16%	46.16%	42.86%	83.33%
	53.28	76.92%	61.54%	46.16%	46.16%	42.86%	83.33%
accuracy	29.60	94.59%	97.30%	97.30%	94.59%	93.24%	98.65%
	35.52	94.59%	97.30%	95.95%	94.59%	93.24%	98.65%
	41.44	95.95%	97.30%	95.95%	94.59%	93.24%	98.65%
	47.36	95.95%	95.95%	94.59%	94.59%	93.24%	98.65%
	53.28	97.30%	95.95%	94.59%	94.59%	93.24%	98.65%

Table 4. Detection accuracy on dataset WBCD.

Metrics	Methods *min_wsup*	MWRPM-Outlier	WMFP-Outlier	WFP-Outlier	Adaptive-KD	LODA	ADWD
precision	273.20	78.24%	82.01%	79.50%	58.16%	53.56%	89.96%
	286.86	79.92%	79.92%	76.99%	58.16%	53.56%	89.54%
	300.52	81.17%	78.24%	74.90%	58.16%	53.56%	90.38%
	314.18	83.26%	76.15%	73.22%	58.16%	53.56%	89.96%
	327.84	84.52%	75.31%	71.97%	58.16%	53.56%	90.79%
recall	273.20	66.31%	77.78%	75.10%	54.30%	47.94%	88.11%
	286.86	69.45%	75.49%	70.77%	54.30%	47.94%	85.94%
	300.52	72.66%	72.76%	67.29%	54.30%	47.94%	87.10%
	314.18	77.13%	70.00%	64.10%	54.30%	47.94%	88.48%
	327.84	79.84%	67.67%	61.87%	54.30%	47.94%	86.45%
F1-measure	273.20	71.78%	79.84%	77.24%	56.16%	50.59%	89.03%
	286.86	74.32%	77.64%	73.75%	56.16%	50.59%	87.70%
	300.52	76.68%	75.40%	70.89%	56.16%	50.59%	88.71%
	314.18	80.08%	72.95%	68.36%	56.16%	50.59%	89.21%
	327.84	82.11%	71.29%	66.54%	56.16%	50.59%	88.57%
accuracy	273.20	72.18%	83.60%	81.55%	65.74%	59.30%	91.51%
	286.86	75.40%	81.84%	77.75%	65.74%	59.30%	89.75%
	300.52	78.62%	79.50%	74.52%	65.74%	59.30%	90.63%
	314.18	82.72%	77.16%	71.30%	65.74%	59.30%	91.80%
	327.84	85.07%	74.82%	68.96%	65.74%	59.30%	90.04%

Table 5. Detection accuracy on dataset ForestCover.

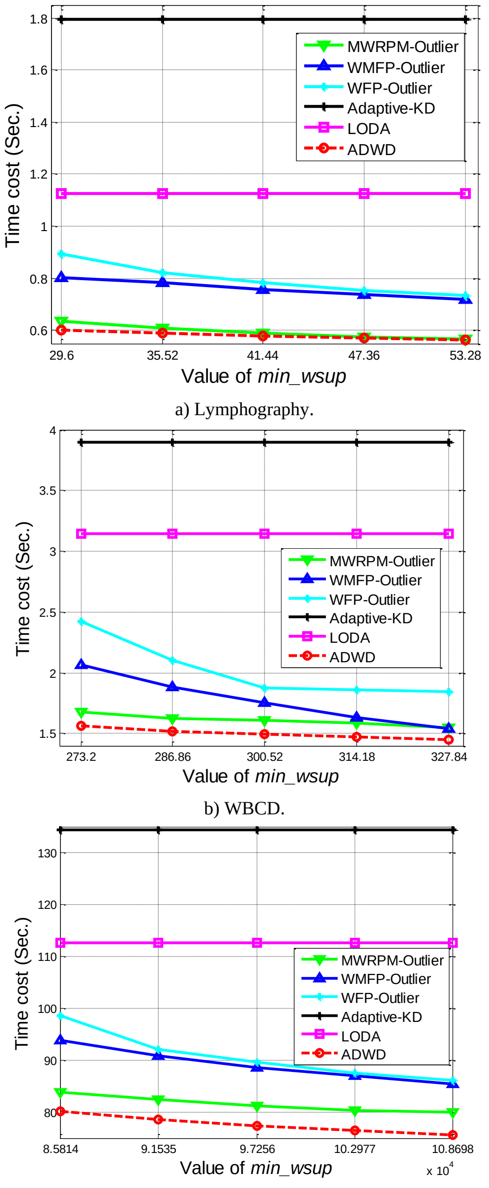
Metrics	Methods *min_wsup*	MWRPM-Outlier	WMFP-Outlier	WFP-Outlier	Adaptive-KD	LODA	ADWD
precision	85814.40	81.47%	82.96%	81.83%	75.28%	74.88%	88.53%
	91535.36	82.31%	82.05%	80.71%	75.28%	74.88%	88.64%
	97256.32	82.82%	81.22%	80.12%	75.28%	74.88%	88.68%
	102977.28	83.76%	80.31%	79.21%	75.28%	74.88%	88.57%
	108698.24	84.64%	79.61%	78.19%	75.28%	74.88%	88.50%
recall	85814.40	77.52%	78.91%	77.62%	72.11%	70.11%	88.34%
	91535.36	79.19%	78.13%	76.13%	72.11%	70.11%	88.26%
	97256.32	80.65%	77.44%	74.21%	72.11%	70.11%	88.20%
	102977.28	81.86%	76.46%	73.61%	72.11%	70.11%	88.41%
	108698.24	83.51%	75.60%	72.03%	72.11%	70.11%	88.37%
F1-measure	85814.40	79.45%	80.88%	79.67%	73.66%	72.42%	88.43%
	91535.36	80.72%	80.04%	78.35%	73.66%	72.42%	88.45%
	97256.32	81.72%	79.28%	77.05%	73.66%	72.42%	88.44%
	102977.28	82.80%	78.34%	76.31%	73.66%	72.42%	88.49%
	108698.24	84.07%	77.55%	74.98%	73.66%	72.42%	88.43%
accuracy	85814.40	99.55%	99.57%	99.55%	99.44%	99.39%	99.78%
	91535.36	99.58%	99.56%	99.51%	99.44%	99.39%	99.77%
	97256.32	99.62%	99.55%	99.47%	99.44%	99.39%	99.77%
	102977.28	99.64%	99.53%	99.45%	99.44%	99.39%	99.78%
	108698.24	99.68%	99.51%	99.42%	99.44%	99.39%	99.78%

As is shown in Table 3 to Table 5 that on the datasets Lymphography, WBCD and ForestCover, the used four evaluation metrics (including precision, recall, F1-measure and accuracy) of the proposed ADWD method are the highest compared with five state-of-the-art methods under these min_wsup values. For the Adaptive-AD and LODA methods, their four evaluation metrics keep constant no matter the change of min_wsup values, it is owing to that the foundation of these two methods is the distance between each data sample and the distribution of each data sample, respectively, which is not influenced by the min_wsup values. Because Adaptive-AD and LODA methods do not consider the weight value of each data sample in the detection of anomalies, thus, their detection accuracy is lower than that of other four anomaly detection methods. For the compared feature analysis-based anomaly detection methods (including MWRPM-Outlier, WMFP-Outlier and WFP-Outlier), their four metrics are changed with the change of min_wsup values. As the min_wsup value is increasing, the precision, recall, F1-measure and accuracy of MWRPM-Outlier show an obviously increase trend, while the WMFP-Outlier and WFP-Outlier are opposite. It is attributed by that in the MWRPM-Outlier method, the MWRIs are used in the determination of anomalies, while the total number of mined MWRIs is much more under large min_wsup values, thus, more associated features can be used in the detection phase. However, the total number of mined MWFIs and weighted frequent itemsets of WMFP-Outlier and WFP-Outlier is much less when the min_wsup value is becoming larger, which results in the decrease of precision, recall, F1-measure and accuracy of these methods. Compared with MWRPM-Outlier, WMFP-Outlier and WFP-Outlier methods, both MWRIs and MWFIs are used in the judgement of outliers, which causes the detection accuracy of proposed ADWD method do not rely on the set of min_wsup values.

5.3. Answer to RQ2

To answer RQ2, we compare the proposed ADWD method with other five state-of-the-arts methods. Each experiment is conducted for 50 times, and the average time cost is calculated and shown in Figure 2.

It is observed from Figure 2 that the time cost of ADWD method is shorter than that of other five compared state-of-the-art methods, especially much shorter than Adaptive-KD and LODA methods, while slightly shorter than MWRPM-Outlier method. The reason for consuming shorter time of ADWD method is that the proposed ADWD method uses tree structure in the associated features mining phase to quickly find different features and thus calculating the weight support of current feature; In addition, the weighted rare 1-itemsets are discarded before constructing the tree structures to reduce the scale of tree structure, and the weighted frequent 1-itemsets are arranged with their decrease weight support to reduce the different parts between extended itemsets. With the above strategies, the MWFIs and MWRIs can be mined from datasets with less time cost. In the compared methods, the time cost of MWRPM-Outlier method is closer to that of ADWD method, which is benefit from only the different itemsets are considered in the calculation of weight support values, which can reduce the computational scales. However, the time cost of Adaptive-KD method is much longer than that of other compared methods, it is attributed by that Adaptive-KD needs to calculate the distance between each data sample, which is time consuming. For the four compared feature analysis-based anomaly detection methods, their time cost shows a decrease trend accompanied with the increase of min_wsup values, it is owing to that more itemsets become WRIs under large min_wsup values and do not participate in the following “itemset expanding” operations.



a) Lymphography.

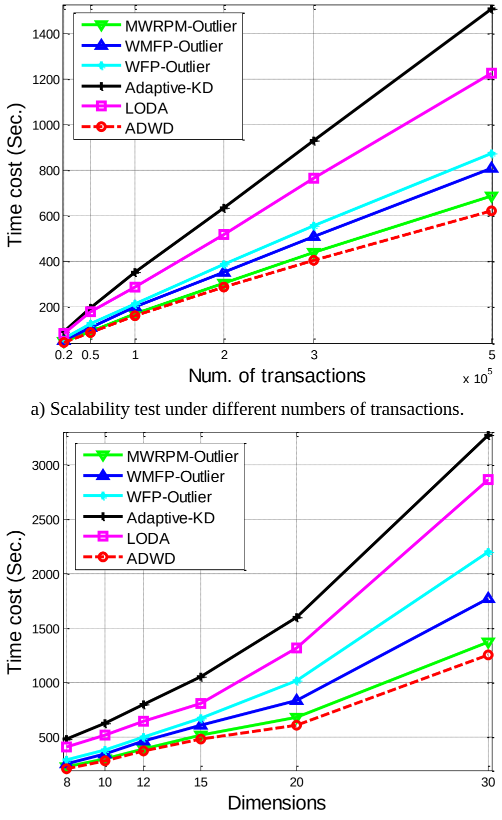
b) WBCD.

c) Forest cover.

Figure 2. Time cost on datasets Lymphography, WBCD and ForestCover.

5.4. Answer to RQ3

To answer RQ3, we use a synthetic dataset generated with the way provided in [5] to test the scalability, where the influence of different numbers of transactions and different dimensions of transactions to the scalability are considered in the experiments. Firstly, the dimension of each transaction is set to 10 and kept constant, and the number of generated transactions is set to 20000, 50000, 100000, 200000, 300000, and 500000, respectively; Secondly, the number of generated transactions is set to 200000 and kept constant, and the dimension of each transaction is set to 8, 10, 12, 15, 20, and 30, respectively. To reduce the randomness of the time cost, each experiment is conducted for 50 times, the average time cost is calculated and shown in Figure 3.



a) Scalability test under different numbers of transactions.

b) Scalability test under different dimensions of transactions.

Figure 3. Scalability test of the compared anomaly detection methods.

It can be known from Figure 3 that with the increase number of transactions and the increase dimensions of transactions, the time cost of ADWD method is the lowest, while the time cost of Adaptive-KD method is the longest, which is very similar to the experimental results on three public datasets. In the compared methods, the time cost of MWRPM-Outlier method is much closer to that of ADWD method, it is owing to that MWRPM-Outlier also uses the different parts of itemsets to reduce the time used in the calculation of min_wsup values, which is verified very useful. With the number of transactions increases, the time cost of six compared methods shows a liner increase trend, while the time cost of six compared methods shows a quadratic trend with the increase dimensions of transactions. The experimental results on the scalability test verify that the proposed ADWD method has a better scalability than that of other five compared methods, it can be used to detect anomalies from high-dimensional datasets and large-scale datasets.

6. Conclusions

To effectively detect the potential anomalies in weighted data and make the anomaly detection result not so dependent on the mined *MWFIs* or *MWRIs*, this paper proposes an anomaly detection method called ADWD for weighted data based on the analysis of feature association. With the mining of *MWFIs* and *MWRIs* from weighted data like the manner of *FP-Growth* method, three deviation metrics are defined to calculate transaction's abnormal score, and then the transactions whose deviation degree in top ranked are judged as anomalies. Massive results on three weight datasets show that ADWD method can obtain high detection accuracy (including precision, recall, F1-measure and accuracy) in the detection of anomalies within less time consumption, as well as has better scalability.

In the future, we would like to use more weighted datasets to verify the efficiency of the proposed ADWD method. In addition, we also would like to consider more influencing factors to further improve detection accuracy for weighted data.

Acknowledgements

This work was partly supported by the Shanxi Agricultural University Youth Science and Technology Innovation Fund (Grant no. 2019017).

References

[1] Agrawal R. and Srikant R., "Fast Algorithms for Mining Association Rules in Large Databases," *in Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499, 1994.

[2] Bhattacharjee P., Garg A., and Mitra P., "KAGO: An Approximate Adaptive Grid-Based Outlier Detection Approach Using Kernel Density Estimate," *Pattern Analysis and Applications*, vol. 24, no. 4, pp. 1825-1846, 2021. https://doi.org/10.1007/s10044-021-00998-6

[3] Bigdeli E., Mohammadi M., Raahemi B., and Matwin S., "Incremental Anomaly Detection Using Two-Layer Cluster-Based Structure," *Information Sciences*, vol. 429, pp. 315-331, 2018. https://doi.org/10.1016/j.ins.2017.11.023

[4] Cai S., Chen J., Chen H., Zhang C., Li Q., Sosu R., and Yin S., "An Efficient Anomaly Detection Method for Uncertain Data Based on Minimal Rare Patterns with the Consideration of Anti-

Monotonic Constraints," *Information Sciences*, vol. 580, pp. 620-642, 2021. https://doi.org/10.1016/j.ins.2021.08.097

[5] Cai S., Sun R., Hao S., Li S., and Yuan G., "An Efficient Outlier Detection Approach on Weighted Data Stream Based on Minimal Rare Pattern Mining," *China Communications*, vol. 16, no. 10, pp. 83-99, 2019. https://doi.org/10.23919/JCC.2019.10.006

[6] Cai S., Li Q., Li S., Yuan G., and Sun, R., "WMFP-Outlier: An Efficient Maximal Frequent-Pattern-Based Outlier Detection Approach for Weighted Data Streams," *Information Technology and Control*, vol. 48, no. 4, pp. 505-521, 2019. https://doi.org/10.5755/j01.itc.48.4.22176

[7] Cai S., Li L., Chen J., Zhao K., Yuan G., Sun R., Sosu, R., and Huang L., "MWFP-Outlier: Maximal Weighted Frequent-Pattern-Based Approach for Detecting Outliers from Uncertain Weighted Data Streams," *Information Sciences*, vol. 591, pp. 195-225, 2022. https://doi.org/10.1016/j.ins.2022.01.028

[8] Cai S., Huang R., Chen J., Zhang C., Liu B., Yin S., and Geng Y., "An Efficient Outlier Detection Method for Data Streams Based on Closed Frequent Patterns by Considering Anti-Monotonic Constraints," *Information Sciences*, vol. 555, pp. 125-146, 2021. https://doi.org/10.1016/j.ins.2020.12.050

[9] Eom S., Oh B., Shin S., and Lee K., "Multi-Task Learning for Spatial Events Prediction from Social Data," *Information Sciences*, vol. 581, pp. 278-290, 2021. https://doi.org/10.1016/j.ins.2021.09.049

[10] Giacometti A. and Soulet A., "Frequent Pattern Outlier Detection Without Exhaustive Mining," *in Proceedings of the 20th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Auckland, pp. 196-207, 2016. https://doi.org/10.1007/978-3-319-31750-2_16

[11] Han J., Pei J., and Yin Y., "Mining Frequent Patterns without Candidate Generation," *in Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, pp. 1-12, 2000. https://doi.org/10.1145/335191.335372

[12] Hemalatha C., Vaidehi V., and Lakshmi R., "Minimal Infrequent Pattern Based Approach for Mining Outliers in Data Streams," *Expert Systems with Applications*, vol. 42, pp. 1998-2012, 2015. https://doi.org/10.1016/j.eswa.2014.09.053

[13] Huo W., Wang W., and Li W., "AnomalyDetect: An Online Distance-Based Anomaly Detection Algorithm," *in Proceedings of 26th International Conference on Web Services*, San Diego, pp. 63-79, 2019. https://doi.org/10.1007/978-3-030-23499-7_5

[14] Jain P., Bajpai M., and Pamula R., "A modified DBSCAN Algorithm for Anomaly Detection in Time-Series Data with Seasonality," *The International Arab Journal of Information Technology*, vol. 19, no. 1, pp. 23-28, 2022. https://doi.org/10.34028/iajit/19/1/3

[15] Ju H., Lee D., Hwang J., Namkung J., and Yu H., "PUMAD: PU Metric Learning for Anomaly Detection," *Information Sciences*, vol. 523, pp. 167-183, 2020. https://doi.org/10.1016/j.ins.2020.03.021

[16] Kou Y., Lu C., and Chen D., "Spatial Weighted Outlier Detection," *in Proceedings of SIAM International Conference on Data Mining*, Philadelphia, pp. 614-618, 2006. https://doi.org/10.1137/1.9781611972764.71

[17] Li G. and Jung J., "Deep Learning for Anomaly Detection in Multivariate Time Series: Approaches, Applications, and Challenges," *Information Fusion*, vol. 91, pp. 93-102, 2023. https://doi.org/10.1016/j.inffus.2022.10.008

[18] Luo Z., He K., and Yu Z., "A Robust Unsupervised Anomaly Detection Framework," *Applied Intelligence*, vol. 52, no. 6, pp. 6022-6036, 2021. https://doi.org/10.1007/s10489-021-02736-1

[19] Rasheed F. and Alhajj R., "A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences," *IEEE Transactions on Cybernetics*, vol. 44, no. 5, pp. 569-582, 2014. https://doi.org/10.1109/TSMCC.2013.2261984

[20] Rezaei F. and Yazdi M., "A New Semantic and Statistical Distance-Based Anomaly Detection in Crowd Video Surveillance," *Wireless Communication and Mobile Computing*, vol. 2021, pp. 5513582, 2021. https://doi.org/10.1155/2021/5513582

[21] Sharma K. and Seal A., "Outlier-Robust Multi-View Clustering for Uncertain Data," *Knowledge-Based Systems*, vol. 211, pp. 106567, 2021. https://doi.org/10.1016/j.knosys.2020.106567

[22] Shi P., Zhao Z., Zhong H., Shen H., and Ding L., "An Improved Agglomerative Hierarchical Clustering Anomaly Detection Method for Scientific Data," *Concurrency and Computation-Practice and Experience*, vol. 33, no. 6, pp. e6077, 2020. https://doi.org/10.1002/cpe.6077

[23] Smrithy G. and Balakrishnan R., "A Statistical-Based Light-Weight Anomaly Detection Framework for Wireless Body Area Networks," *Computer Journal*, vol. 65, no. 7, pp. 1752-1759, 2022. https://doi.org/10.1093/comjnl/bxab016

[24] Wang W. and Sun D., "The Improved Adaboost Algorithms for Imbalanced Data Classification," *Information Sciences*, vol. 563, pp. 358-374, 2021. https://doi.org/10.1016/j.ins.2021.03.042

[25] Yuan G., Cai S., and Hao S., "A Novel Weighted Frequent Pattern-Based Outlier Detection Method Applied to Data Stream," *in Proceedings of the IEEE 4th International Conference on*

Cloud Computing and Big Data Analysis, Chengdu, China, pp. 503-510, 2019. https://doi.org/10.1109/ICCCBDA.2019.8725699

[26] Zeng S., Zhang B., Gou J., Xu Y., and Huang W., "Fast and Robust Dictionary-based Classification for Image Data," *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 6, pp. 1-22, 2021. https://doi.org/10.1145/3449360

[27] Zhang L., Zhao J., and Li W., "Online and Unsupervised Anomaly Detection for Streaming Data Using an Array of Sliding Windows and PDDs," *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 2284-2289, 2021. https://doi.org/10.1109/TCYB.2019.2935066

Jiayao Li is a lecturer in Software College, Shanxi Agricultural University, China. She received the MS degree from China Agricultural University, China, in 2018. Her major research interests include Multi-Modal Learning, Data Processing, Outlier Detecting.

Rui Wang is a lecturer in Shanxi Polytechnic College, China. She received the MS degree from Northwest A&F University, China, in 2017. Her major research includes Data Processing, Outlier Detecting, Image Recognition Technology.